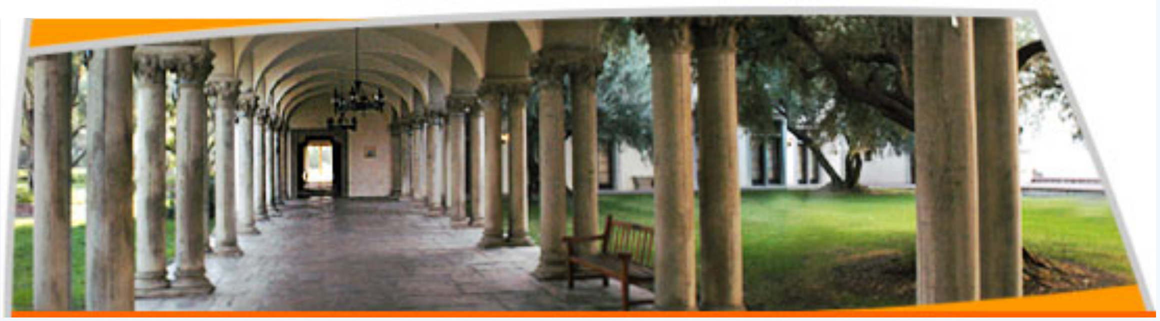
National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Facilitating Distributed Climate Modeling Research and Analysis via the Climate Data eXchange



Dan Crichton
Chris Mattmann
Amy Braverman

18 September 2008
GO ESSP 2008 Workshop

NASA's Satellite Data and Climate Research

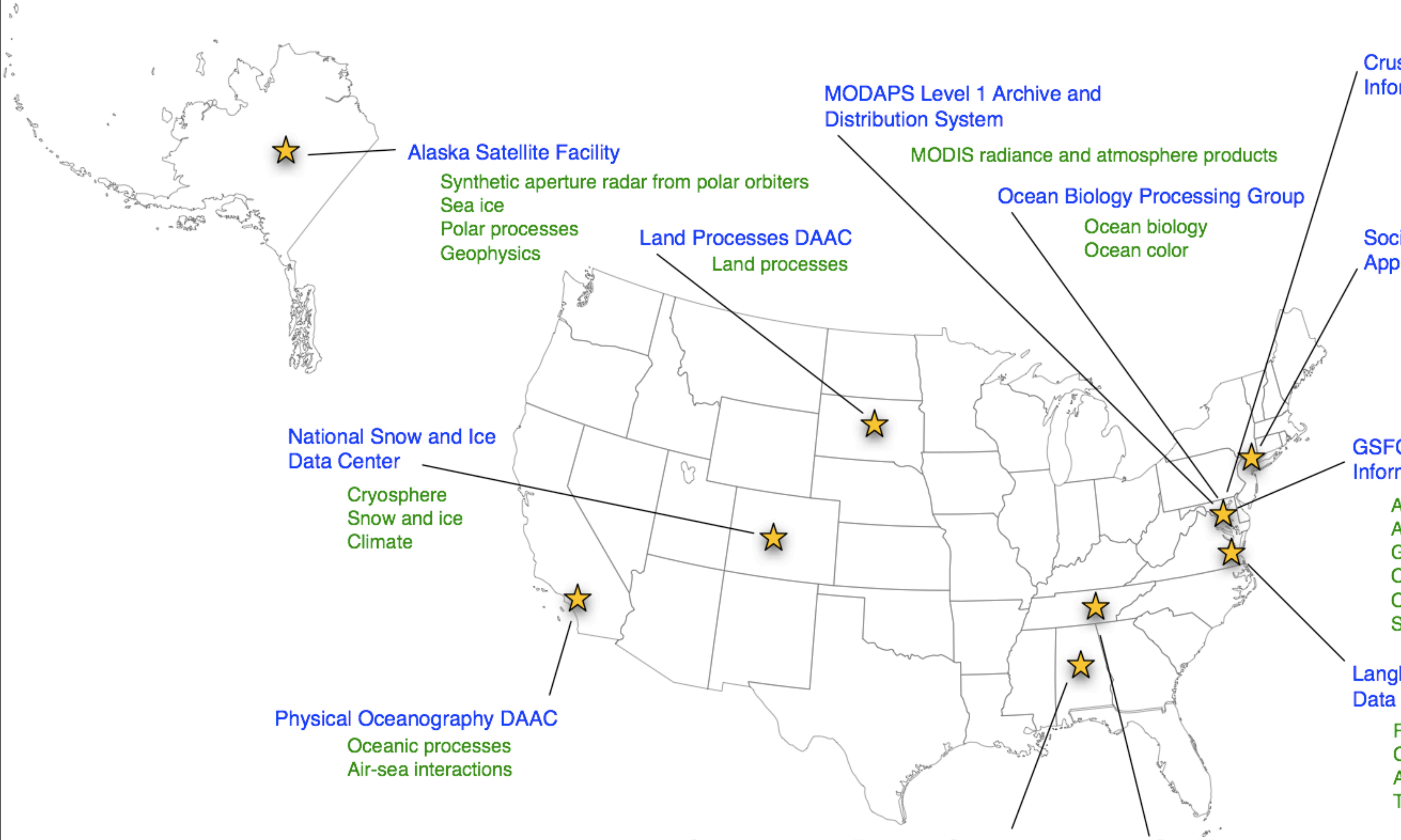
- **Two major legacies from NASA's Earth Observing System Data and Information System (EOSDIS)**
  - *Archiving of explosion in observational data in Distributed Active Archive Centers (DAACs)*
    - Request-driven retrieval from archive is time consuming
  - *Adoption of Hierarchical Data Format (HDF) for data files*
    - Defined by and unique to each instrument but not necessarily consistent between instruments

- **What are the next steps to accelerating use of an ever increasing observational data collection?**
  - *What data are available?*
  - *What is the information content?*
  - *How should it be interpreted in climate modeling research?*

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

EOSDIS DAAC’s
Earth Observing System Data and Information System
Distributed Active Archive Centers



Alaska Satellite Facility
- Synthetic aperture radar from polar orbiters
- Sea ice
- Polar processes
- Geophysics

Land Processes DAAC
- Land processes

MODAPS Level 1 Archive and Distribution System
- MODIS radiance and atmosphere products

Ocean Biology Processing Group
- Ocean biology
- Ocean color

Crustal Dynamics Data Information System
- Crustal dynamics

Socioeconomic Data and Applications Data Center
- Population
- Sustainability
- Geospatial data
- Multilateral environmental agreements

National Snow and Ice Data Center
- Cryosphere
- Snow and ice
- Climate

GSFC Earth Science Data and Information Services Center
- Atmospheric composition
- Atmospheric dynamics
- Global precipitation
- Ocean biology
- Ocean dynamics
- Solar irradiance

Physical Oceanography DAAC
- Oceanic processes
- Air-sea interactions

Langley Atmospheric Sciences Data Center
- Radiation budget
- Clouds
- Aerosol
- Tropospheric chemistry

Global Hydrology Resource Center
- Hydrological cycle
- Severe weather interactions
- Lightening
- Convection

Oak Ridge National Laboratory Data Center
- Biogeochemical dynamics
- Ecology
- Environmental processes

Data Processing Levels

Level 0
Reconstructed, unprocessed instrument/payload data at full resolution; any and all communications artifacts, e.g., synchronization frames, communications headers, duplicate data removed.

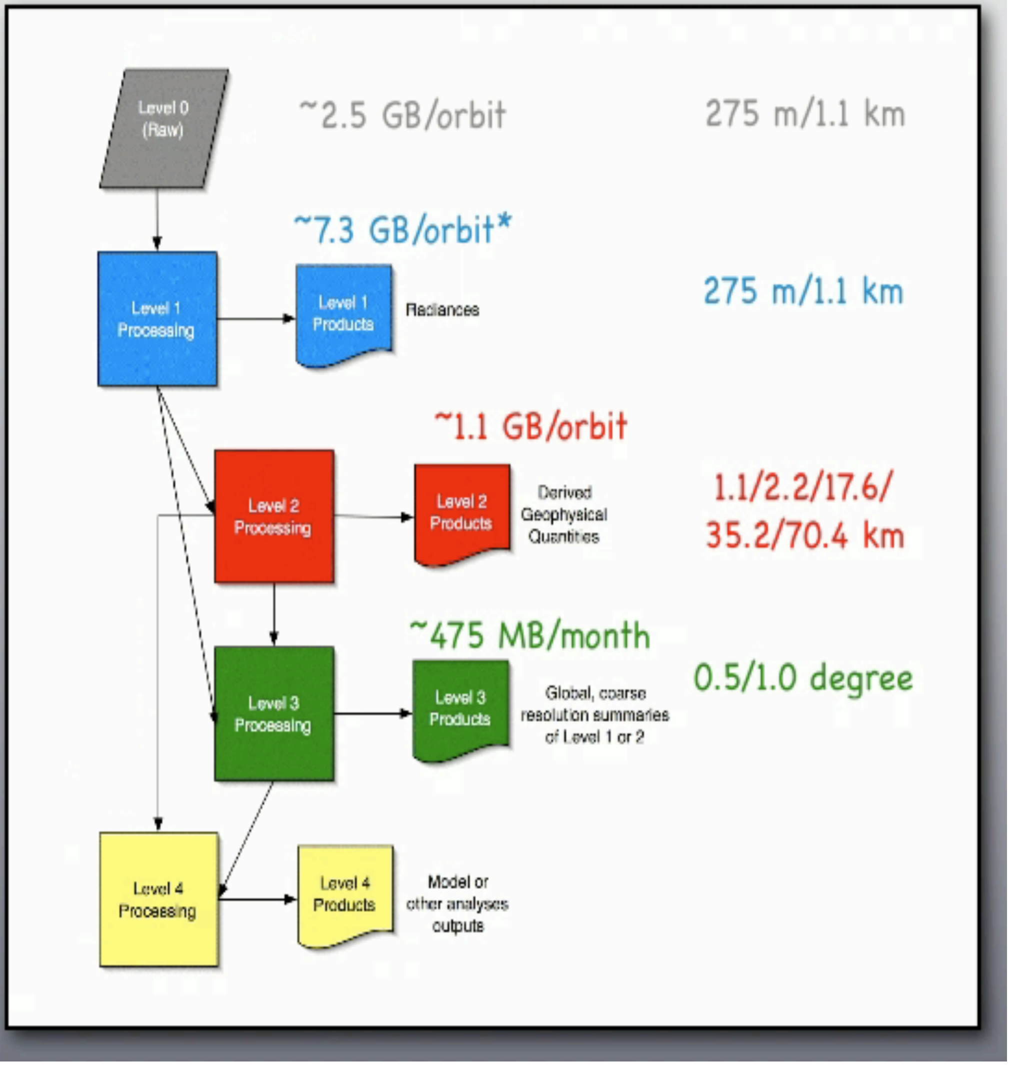
Level 1A
Reconstructed, unprocessed instrument data at full resolution, time-referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters, e.g., platform ephemeris, computed and appended but not applied to the Level 0 data.

Level 1B
Level 1A data that have been processed to sensor units (not all instruments will have a Level 1B equivalent).

Level 2
Derived geophysical variables at the same resolution and location as the Level 1 source data.

Level 3
Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.

Level 4
Model output or results from analyses of lower level data, e.g., variables derived from multiple measurements.

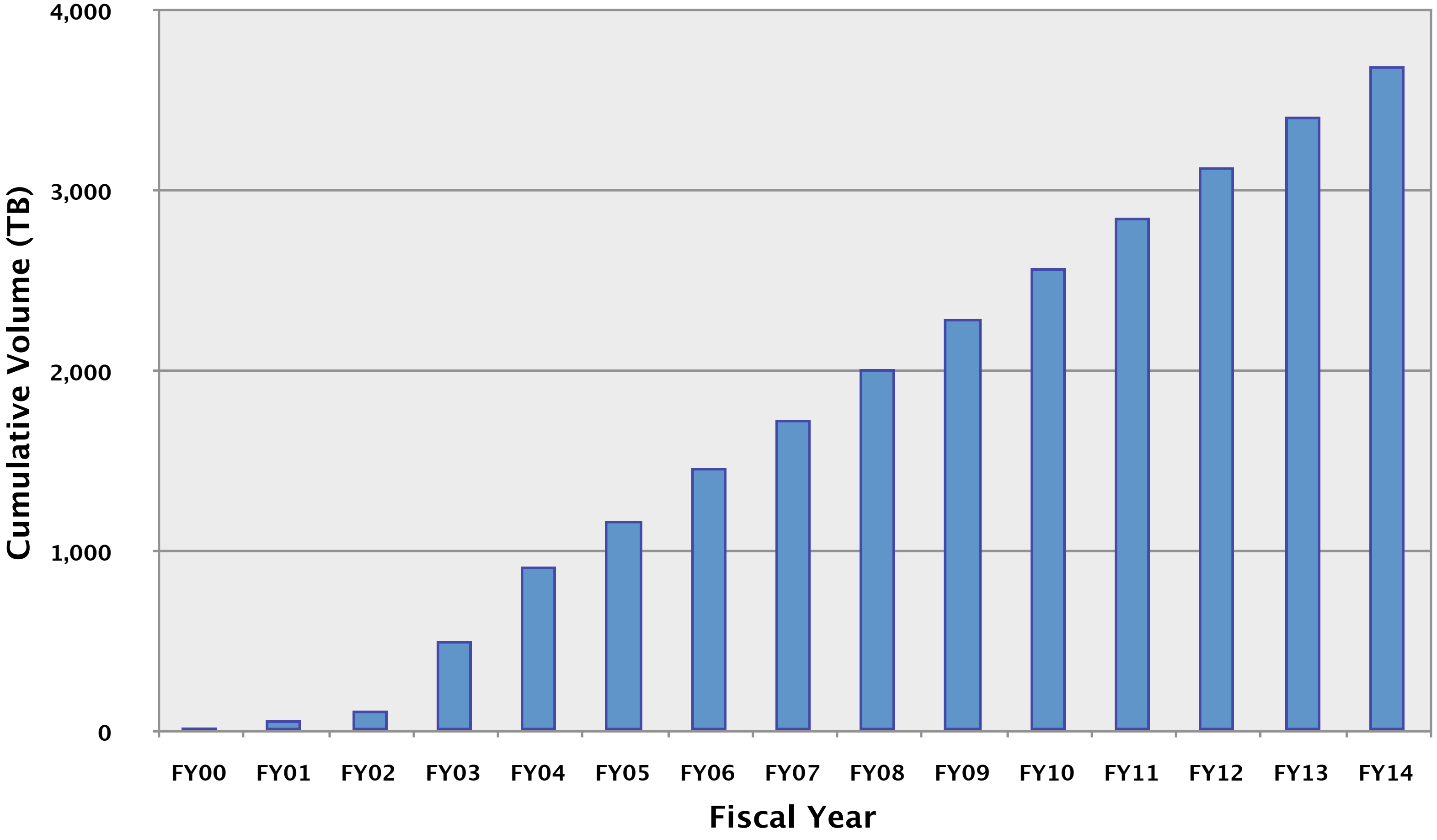


National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

EOSDIS DAAC's
Earth Observing System Data and Information System
Distributed Active Archive Centers

Cumulative Volume of L2+ Products at All DAACs



Researcher’s Challenge

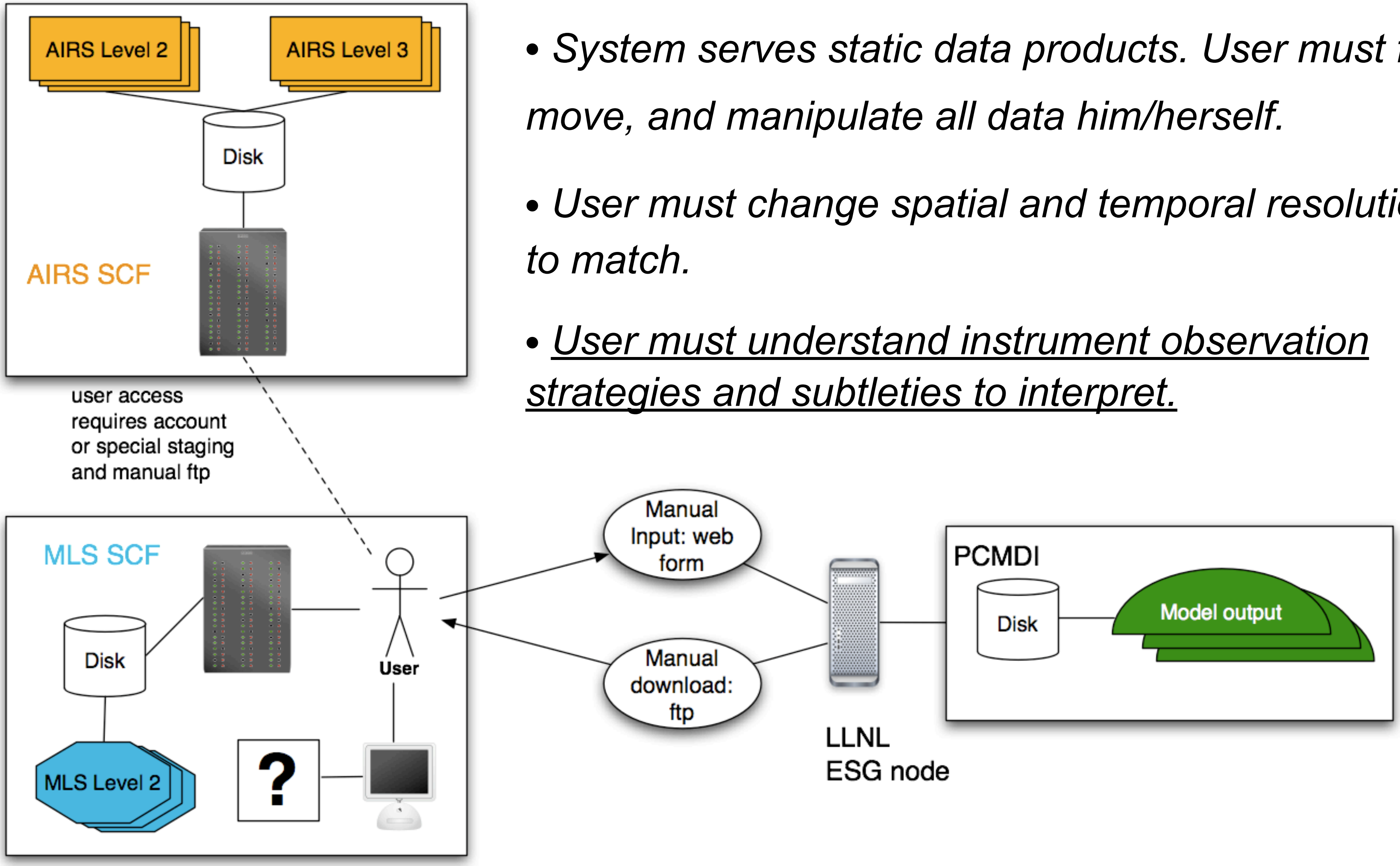
- **Scientists cannot easily locate, access, or manipulate observational data or model output necessary to support climate research**
  - *The latest data are available from independent instrument project data systems.*
  - *Scientists may not even be aware of what repositories or data exist*
  - *Observational data and model output data are heterogeneous in form and cannot be simply compared or combined.*

- **Research data systems are often ad-hoc**
  - *They lack a modular approach limiting extensibility*
  - *They are designed individually rather than as a system*
  - *There are few capabilities in common between systems*

- **They require “human-in-the-loop”**
  - *Web forms, manual ftp transfer*
  - *Rectification left to individual scientists*

Current Data System

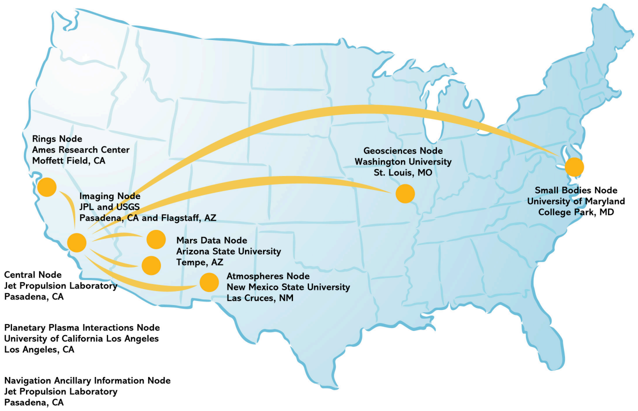
- *System serves static data products. User must find move, and manipulate all data him/herself.*
- *User must change spatial and temporal resolutions to match.*
- *User must understand instrument observation strategies and subtleties to interpret.*



Experience in Planetary Science: NASA's PDS

- Pre-Oct 2002, no unified view across distributed operational planetary science data repositories
  - Science data distributed across the country
  - Science data distributed on physical media
- Planetary data archive increasing from 4 TBs in 2001 to 100+ TBs in 2008
  - Traditional distribution infeasible due to cost and system constraints
  - Mars Odyssey could not be distributed using traditional method
- PDS now has a distributed, federated framework in place
  - Support online distribution of science data to planetary scientists
  - Enable interoperability between nine institutions
  - Support real-time access to distributed catalogs and repositories
  - Uniform software interfaces to all PDS data holdings scientists and developers to link in their own tools
  - Moving towards international standardization with the International Planetary Data Alliance
  - Operational October 1, 2002

2001 Mars Odyssey



PDS Federation

Experience in Cancer Research: NCI's EDRN

- Experience in science information systems has lead to interagency agreements with both NIH and NCI
- Provided the NCI with a bioinformatics infrastructure for establishing a virtual knowledge system
  - Currently deployed at 15 of 31 NCI Research Institutions for the Early Detection Research Network (EDRN)
  - Providing real-time access to distributed, heterogeneous databases
  - Capturing validation study results, instrument results images, biomarkers, protocols, etc
  - Funded 2001-2010 for NCI's Early Detection Research Network
- Currently working with a new initiative in establishing an "informatics plan" for the Clinical Proteomics Technology Initiative

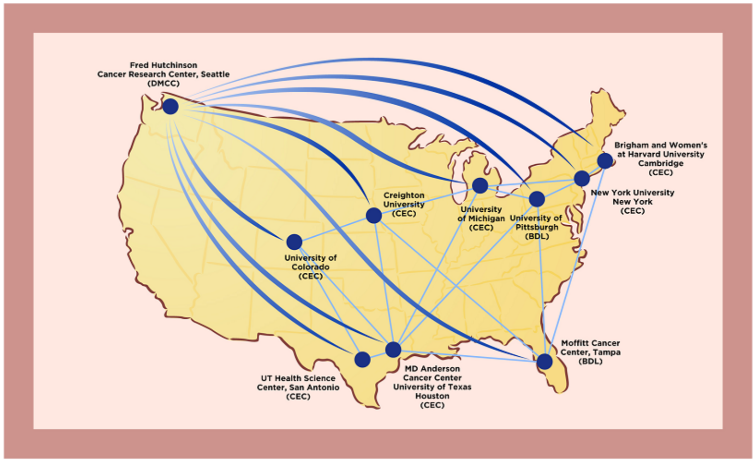
NATIONAL CANCER INSTITUTE®

Cancer Biomarkers Group
Division of Cancer Prevention

Early Detection Research Network

FRED HUTCHINSON CANCER RESEARCH CENTER

Advancing Knowledge, Saving Lives



CDX

- **What: build open source software to**

  -- connect existing systems into a virtual network (big disk),

  -- push as much computation as possible into remote nodes to minimize movement of data,

  -- operators to rectify and fuse heterogeneous data sets, provide uncertainties.

- **Why: scientists need command line access to data sets (model output and observations) such that all data look local and rectified.**

- **How: use technologies in new ways**

  -- distributed computing technologies already in place at JPL (OODT, others); Earth System Grid for parallel transfer,

  -- rigorous mathematical/statistical methods for interpolation, transformation, fusion, and comparisons. Comparisons require new methods developed specifically for massive, distributed data sets. Uncertainties are key.

- **Why is this different:**

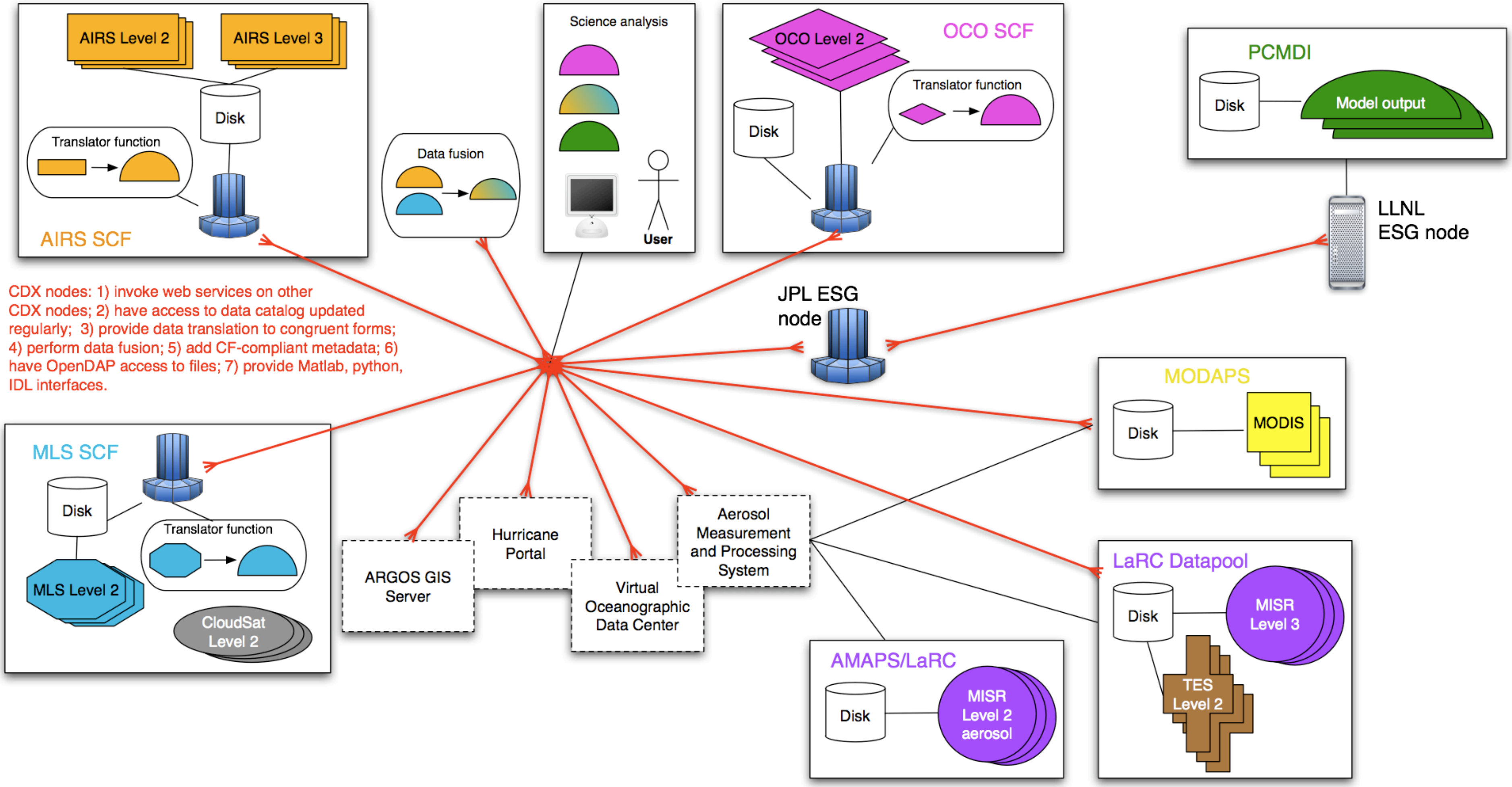
  -- system will capture intellectual capital of instrument scientists and modelers through multiple, flexible operators,

  -- NOT trying to be all things to all people!

National Aeronautics and Space Administration

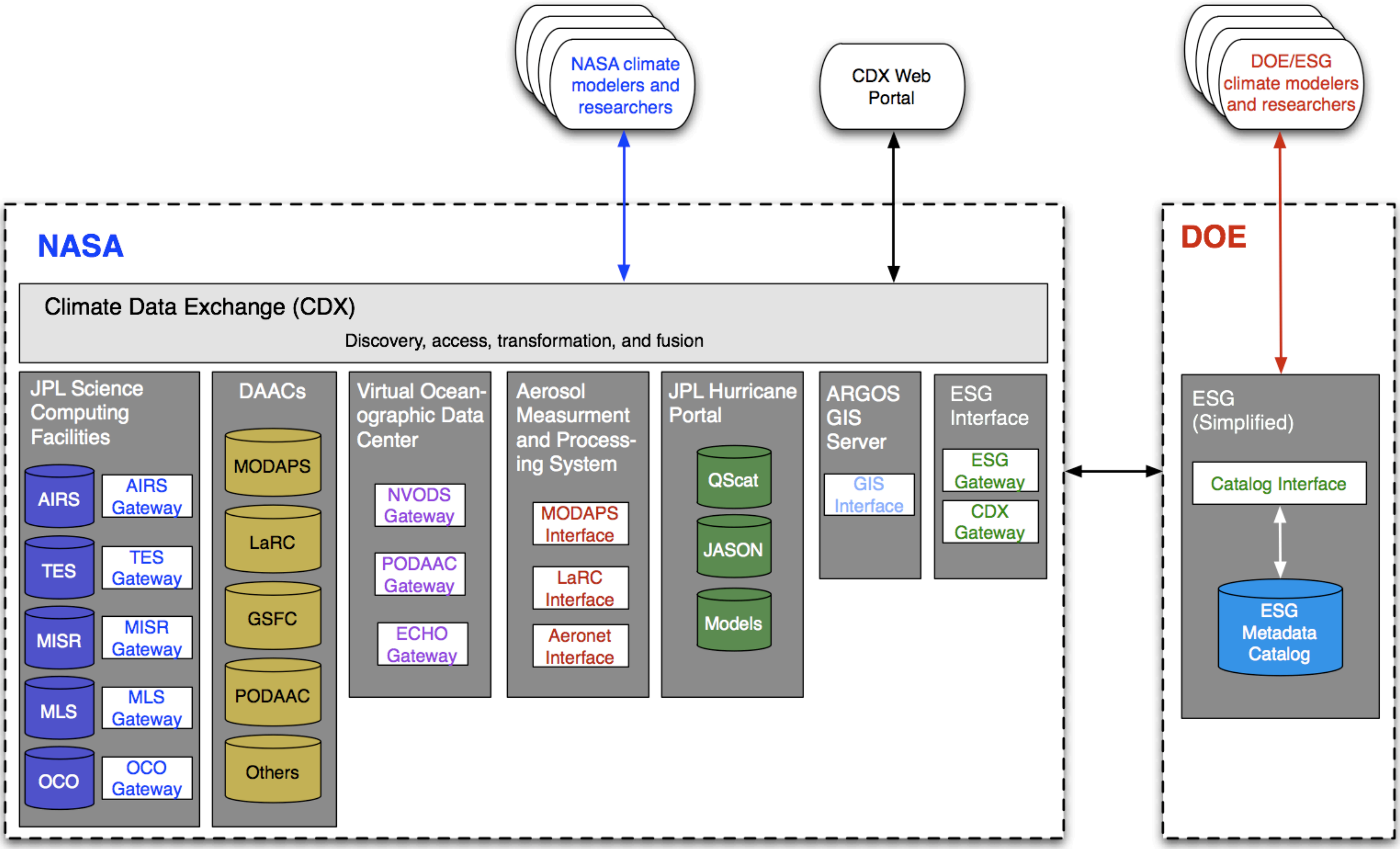
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Climate Data eXchange Research Flow



CDX nodes: 1) invoke web services on other CDX nodes; 2) have access to data catalog updated regularly; 3) provide data translation to congruent forms; 4) perform data fusion; 5) add CF-compliant metadata; 6) have OpenDAP access to files; 7) provide Matlab, python, IDL interfaces.

Climate Data eXchange Architecture



NASA climate modelers and researchers

CDX Web Portal

DOE/ESG climate modelers and researchers

NASA

Climate Data Exchange (CDX)

Discovery, access, transformation, and fusion

JPL Science Computing Facilities

AIRS

AIRS Gateway

TES

TES Gateway

MISR

MISR Gateway

MLS

MLS Gateway

OCO

OCO Gateway

DAACs

MODAPS

LaRC

GSFC

PODAAC

Others

Virtual Oceanographic Data Center

NVODS Gateway

PODAAC Gateway

ECHO Gateway

Aerosol Measurment and Processing System

MODAPS Interface

LaRC Interface

Aeronet Interface

JPL Hurricane Portal

QScat

JASON

Models

ARGOS GIS Server

GIS Interface

ESG Interface

ESG Gateway

CDX Gateway

DOE

ESG (Simplified)

Catalog Interface

ESG Metadata Catalog

Conclusions

- **CDX is a paradigm shift in data access/delivery/analysis systems**
  - *Data analysis should not be decoupled from access and delivery*
  - *Should support interactive analysis*
- **Distributed computing (e.g. web services) architecture is key**
  - *Support remote query, access, and computation*
  - *Not tied to any particular implementation*
  - *ESG is a success story for access and delivery*
  - *Partnership between JPL and LLNL to extend success to interactive, distributed data analysis*
- **JPL will develop, deploy and test V1.0 of CDX over next 18 months**
  - *Funded NASA support to construct JPL ESG data node*
  - *Critical components proposed for internal support at JPL to enable model evaluations, validation, and projections*
- **Feedback, suggestions, and collaborations welcome on path forward**

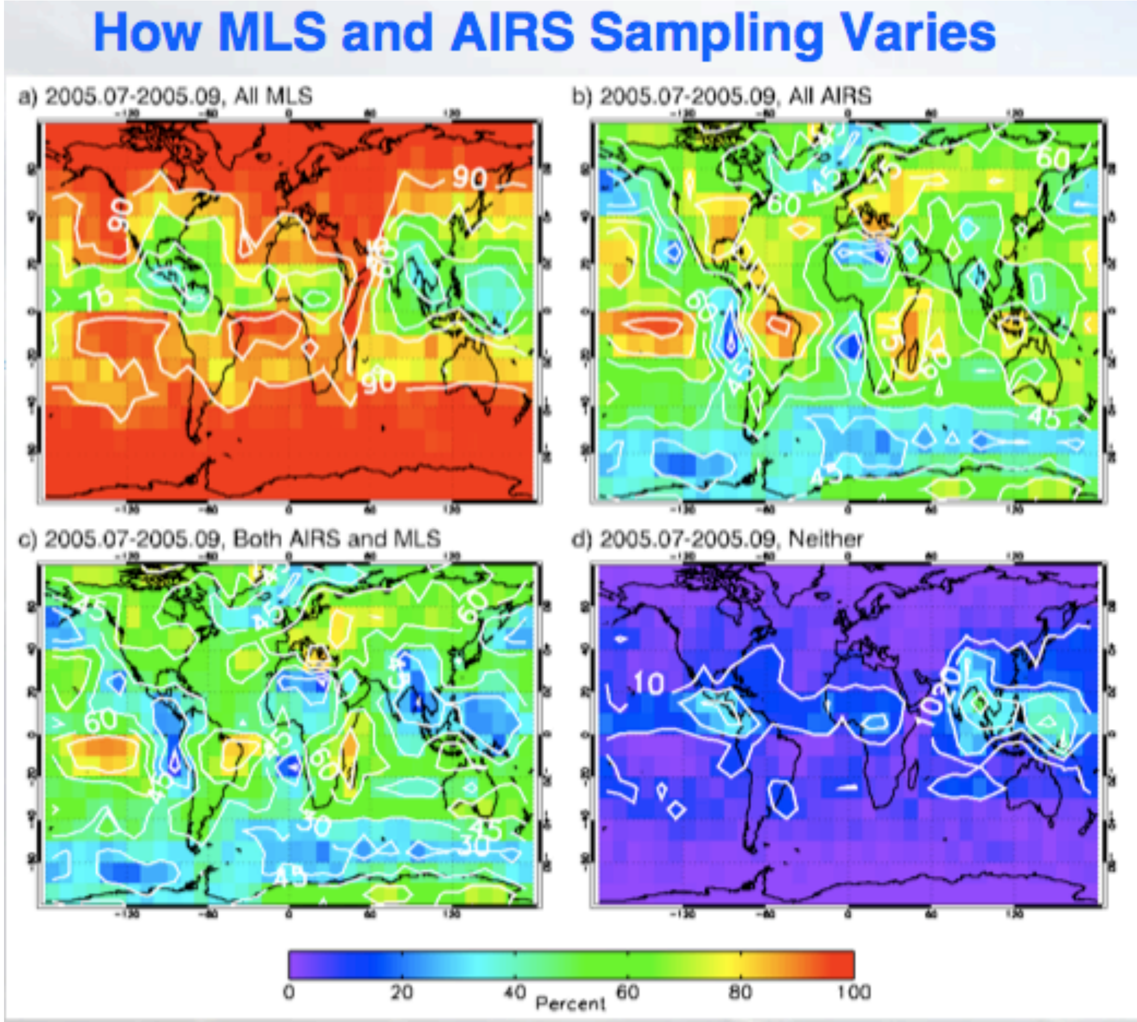
Backup

Climate Research Use Case

- **What is radiative effect of the vertical distribution of water vapor in the atmosphere under clear-sky conditions?**
  - *Warming by water vapor back to the surface could lead to increased evaporation and accelerate (positive feedback) the "greenhouse effect"*

- **Investigation and validation of climate model representations of water vapor distributions can be made by comparison to both AIRS and MLS measurements of water vapor**
  - *AIRS provides water vapor measurements up to 200 mb (15km)*
  - *MLS provides water vapor measurements from 300 mb to 100 mb (8km to 18km)*
  - *AIRS and MLS sample different states: each is capable of measuring vapor in clear scenes, but under cloudy conditions they have different biases.*
  - *Need to combine these data to get the full picture.*

Combining Instrument Data to enable Climate Research: AIRS and MLS


How MLS and AIRS Sampling Varies
a) 2005.07-2005.09, All MLS
b) 2005.07-2005.09, All AIRS
c) 2005.07-2005.09, Both AIRS and MLS
d) 2005.07-2005.09, Neither
90
75
60
45
30
10
0
20
40
60
80
100
Percent

Combining AIRS and MLS requires:

- *Rectifying horizontal, vertical and temporal mismatch*

- *Assessing and correcting for the instruments' scene-specific error characteristics* (see left *diagram)*